DOCUMENT RESUME

ED 269 461 TM 860 315

AUTHOR Cope, Ronald T.
TITLE How Well Do the Angoff Design V Linear Equating Methods Stack Up against the Tucker and Levine Methods?
PUB DATE Apr 86
NOTE 32p.; Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Certification; Comparative Analysis; *Equated Scores; Higher Education; *Mathematical Models; Occupational Tests; Professional Occupations; Raw Scores; Simulation; *Statistical Bias; Statistical Studies; *Test Theory
IDENTIFIERS *Angoff Methods; Levine Equating Method; *Linear Equating Method; Tucker Common Item Equating Method

ABSTRACT

Comparisons were made of three Angoff Design V linear equating methods (two forms equated to a common test, two forms predicted by a common test, or two forms used to predict a common test) and Tucker's and R. Levine's linear methods, under common item linear equating with non-equivalent populations. Forms of a professional certification test administered to over 78,000 applicants were equated under the five different methods, using single-link equating of selected pairs of forms and cyclical equating of selected forms to themselves via equating chains. In addition, the five methods were tried on a set of hypothetical situations involving parallel or congeneric forms and anchor tests. In the single-link equatings, raw score equivalents by the Angoff methods tended to fall between those obtained with the Tucker and Levine methods. The chain equatings produced similar estimated bias and estimated root-mean-squared error of score equivalents for the five different methods. Results of equating in the hypothetical situations appeared consistent with the manner in which they were constructed and with the assumptions underlying the five different equating methods. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED269461

How Well Do the Angoff Design V Linear Equating Methods Stack Up Against the Tucker and Levine Methods?

Ronald T. Cope
The American College Testing Program

RUNNING HEAD: Tucker, Levine, and Angoff Design V Equating

Paper presented at the Annual Meeting of the
American Educational Research Association
San Francisco, April 1986

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. T. Cope

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 860 315

ABSTRACT

Comparisons were made of three Angoff Design V linear equating methods and the Tucker and Levine linear methods under common item linear equating with non-equivalent populations. Forms of a professional certification test were equated under the five different methods using both single-link equating of selected pairs of forms and cyclical equating of selected forms to themselves via equating chains. In addition, the five methods were tried on a set of hypothetical situations involving parallel or congeneric forms and anchor tests. In the single-link equatings, raw score equivalents by the Design V methods tended to fall between those obtained by use of the Tucker and Levine methods. The chain equatings produced similar estimated bias and estimated root-mean-squared error of score equivalents for the five different methods. Results of equating in the hypothetical situations appeared consistent with the manner in which they were constructed and with the assumptions underlying the five different equating methods.

KEY WORDS: Equating, Linear Equating, Tucker Method, Levine Method

Consider these three ways to obtain "equivalent" test scores:

1. Equate Form X to anchor test V; then equate V to Form Y.
2. Find score x_0 on Form X and score y_0 on Form Y, both of which are predicted by a given score v_0 on anchor test V. x_0 and y_0 are the equated scores.
3. Find score x_1 on Form X and score y_1 on Form Y, both of which predict the same score v_0 on anchor test V. x_1 and y_1 are the equated scores.

Forms X and Y and the anchor test are constructed to be similar in content and difficulty level.

In his "Scales, norms, and equivalent scores" chapter, Angoff (1971) outlines these procedures under "Design V: Other methods involving score data." Design V comprises three different methods: (1) Forms X and Y equated to a common test, (2) Forms X and Y predicted by a common test, and (3) Forms X and Y predicting a common test.

For each of these three methods, Angoff outlines both a linear and an equipercentile procedure. In this paper I research only the linear procedures.

In their discussion of observed-score test equating, Braun and Holland (1982) discuss the equipercentile procedure of Design V, Method 1. They show that equipercentile equating of Forms X and Y through a third test V is equivalent to the use of a synthetic population and frequency estimation in the equating of Y to X or X to Y only if certain unusual relationships hold among the density functions of X, Y, and V. They state that "similar results can be given for linear equating."

The purpose of this study is to make both simulated and real-data comparisons, under common-item linear equating, of the often-used Tucker and

Levine methods with the Angoff Design V methods. All five methods use statistics on a set of common items (anchor test) that are answered by both the group taking one form of a test and the group taking another form of the test. In all these methods, anchor test scores are used as a basis for adjusting for differences between the two groups of examinees taking the two different forms. The Design V methods, however, use the anchor test in a less direct manner than the Tucker and Levine methods, even though the variations noted above under 2 and 3 employ the same statistics as the Tucker and Levine methods (variation 1 does not use correlations between the forms and the anchor test).

For detailed discussions of the Levine and Tucker methods see Levine (1956) and Braun and Holland (1982, pp 23-24). See Appendix A for a summary of the equating formulas used for these two methods.

METHOD

Subjects

Examinees used in this study consist of five groups of applicants in a professional certification program. Each one of these five groups took one of five forms of the certification test in a particular administration. All are first-timers (nonrepeaters). Table 1 gives the number of first-time examinees who took each of the five forms, as well as means and standard deviations of their raw scores.

Insert Table 1 about here

Measures

Measures are five forms of a multiple-choice professional certification test with a raw score range of 0 to 280. The test is administered twice a

year at six-month intervals. Forms B, D, and E were administered in the spring; Forms A and C were administered in the fall. For each administration a new form is constructed, used in the administration, and then retired. Except for retired items used in a sample test, no future, current, or past test items are released to examinees or otherwise publicized.

Procedure

The following three ways of comparing the different equating methods are used:

1. Single-link comparisons. Compare results of equating pairs of forms of a professional certification test under the five different linear methods.
2. Equating-chain comparisons (equating forms to themselves). Compare equatings around three cyclical chains of links among different forms of the certification test. These chains are shown in Figure 1.
3. Parametric model comparisons. Assume an ideal world of parallel tests, known parameters (infinite examinee populations), and congeneric or parallel anchor tests. Compare results of applying the five methods to these idealized situations.

Insert Figure 1 about here

Procedure 1 allows us to see where equated raw scores under the three Design V methods stand in relation to those obtained under the Tucker and Levine methods. It permits no comparison of relative accuracy of the five methods because we expect the raw score scales of two forms to differ in an

unknown way. If we knew somehow that the raw score scales of two forms were the same, then there would be no need to try to equate the forms.

Procedure 2 compares the accuracy of the different methods by equating two selected forms to themselves via a circular chain (cycle) of equatings among forms. If an equating method were free of systematic and random error, we would expect the equating function, $y = Ax + B$, to have a slope, A, of 1 and an intercept, B, of 0.

Procedure 3 disregards the element of sampling error of equating by envisioning an ideal world of parallel tests (sometimes approximated in reality) and infinite examinee populations yielding known parameters. An accurate equating method ought to produce for parallel forms an identity equating function ($y = x$) — or at least something close to it. How will the five methods compare in this respect?

Derivation of Angoff Design V Equating Functions

Method V1. Equate Form X to anchor test V; then equate V to Form Y. Set corresponding z-scores on X and V equal to each other,

$$\frac{x - \mu(X)}{\sigma(X)} = \frac{v - \mu_x(V)}{\sigma_x(V)} \quad .$$

Note that $\mu_x(V)$ is the mean on anchor test V of the population taking Form X and $\sigma_x(V)$ is the standard deviation of this population on V. These quantities are similarly defined for the population taking Form Y. Solving for v, we get

$$v = \frac{\sigma_x(V)}{\sigma(X)} \; [x - \mu(X)] + \mu_x(V) \quad . \qquad (1)$$

Now equate V to Y by setting corresponding z-scores equal,

$$\frac{v - \mu_y(V)}{\mu_y(V)} = \frac{y - \mu(Y)}{\mu(Y)} \quad .$$

Solve for v,

$$v = \frac{\sigma_y(V)}{\sigma(Y)} [y - \mu(Y)] + \mu_y(V) \qquad (2)$$

Set the right side of Equation 1 equal to the right side of Equation 2, and solve for X,

$$x = \frac{\sigma(X)\sigma_y(V)}{\sigma(Y)\sigma_x(V)} \quad y + \mu(X) - \frac{\sigma(X)\sigma_y(V)}{\sigma(Y)\sigma_x(V)} \mu(Y) + \frac{\sigma(X)}{\sigma_x(V)} \mu_y(V) - \frac{\sigma(X)}{\sigma_x(V)} \mu_x(V).$$

Note that $\frac{\sigma(X)\sigma_y(V)}{\sigma(Y)\sigma_x(V)}$ gives the slope of the Method V1 linear equating function and that the remaining four terms give the intercept. In actual equating, one must substitute estimates of the various parameters.

Method V2. Forms X and Y predicted by a common test.

The derivation of the linear equating function for this method is also easy. For this method and method V3, correlations of the anchor test V with Forms X and Y come into play.

First consider linear prediction of Form X scores from scores on anchor test V,

$$x = \rho(X,V) \frac{\sigma(X)}{\sigma_x(V)} \quad [v_x - \mu_x(V)] + \mu(X)$$

$$= \rho(X,V) \frac{\sigma(X)}{\sigma_x(V)} v_x - \rho(X,V) \frac{\sigma(X)}{\sigma_x(V)} \mu_x(V) + \mu(X) \quad ,$$

Solve for v_x,

BEST COPY AVAILABLE

$$v_x = \frac{x + \rho(X,V) \frac{\sigma(X)}{\sigma_x(V)} \mu_x(V) - \mu(X)}{\rho(X,V) \frac{\sigma(X)}{\sigma_x(V)}} \quad . \tag{3}$$

Similarly, we can predict Form Y scores from V scores,

$$y = \rho(Y,V) \frac{\sigma(Y)}{\sigma_y(V)} [v_y - \mu_y(V)] + \mu(Y) \quad .$$

Solving for v_y, we get

$$v_y = \frac{y + \rho(Y,V) \frac{\sigma(Y)}{\sigma_y(V)} \mu_y(V) - \mu(Y)}{\rho(Y,V) \frac{\sigma(Y)}{\sigma_y(V)}} \quad . \tag{4}$$

We want to know which X and Y scores are predicted by a given V score, so we need to let $v_x = v_y$. Set the right sides of Equations 3 and 4 equal to each other,

$$\frac{x + \rho(X,V) \frac{\sigma(X)}{\sigma_x(V)} \mu_x(V) - \mu(X)}{\rho(X,V) \frac{\sigma(X)}{\sigma_x(V)}} = \frac{y + \rho(Y,V) \frac{\sigma(Y)}{\sigma_y(V)} \mu_y(V) - \mu(Y)}{\rho(Y,V) \frac{\sigma(Y)}{\sigma_y(V)}}$$

Now solve for x to obtain the equating function,

$$x = \frac{\rho(X,V)\ \sigma(X)\ \sigma_y(V)}{\rho(Y,V)\ \sigma(Y)\ \sigma_x(V)} \quad y + \mu(X)$$

$$+ \frac{\rho(X,V)\ \sigma(X)\ \sigma_y(V)}{\rho(Y,V)\ \sigma(Y)\ \sigma_x(V)} [\rho(Y,V) \frac{\sigma(Y)}{\sigma_y(V)} \mu_y(V) - \mu(Y)]$$

$$- \rho(X,V) \frac{\sigma(X)}{\sigma_x(V)} \mu_x(V) \quad .$$

Method V3. Forms X and Y predicting a common test. Consider predicting V scores from X scores:

$$v_x = \rho(X,V) \frac{\sigma_x(V)}{\sigma(X)} [x - \mu(X)] + \mu(V) \quad . \tag{5}$$

Similarly, we have for predicting V scores from Y scores:

$$v_y = \rho(Y,V) \frac{\sigma_y(V)}{\sigma(Y)} [y - \mu(Y)] + \mu(V) \quad . \tag{6}$$

We want to find X and Y scores that predict the same V score.

Let $v_x = v_y$. Now we can set the right sides of Equations 5 and 6 equal to each other.

$$\rho(X,V) \frac{\sigma_x(V)}{\sigma(X)} [x - \mu(X)] + \mu(V) = \rho(Y,V) \frac{\sigma_y(V)}{\sigma(Y)} [y - \mu(Y)] + \mu(V)$$

Solve for x to obtain

$$x = \frac{\rho(Y,V)\ \sigma(X)\ \sigma_y(V)}{\rho(X,V)\ \sigma(Y)\ \sigma_x(V)}\ y + \mu(X) + \frac{-\mu_x(V) - \rho(Y,V) \frac{\sigma_y(V)}{\sigma(Y)} \mu(Y) + \mu_y(V)}{\rho(X,V) \frac{\sigma_x(V)}{\sigma(X)}}.$$

Please note:

1. If anchor test V correlates perfectly with both Form X and Form Y, then $\rho(X,V) = \rho(Y,V) = 1$, and Method V3 reduces to Method V1.
2. If V correlates equally with X and Y, then $\frac{\rho(Y,V)}{\rho(X,V)} = \frac{\rho(X,V)}{\rho(Y,V)} = 1$ and all three equating functions have the same slope.

RESULTS

1. Single-link equating.

Tables 2, 3, and 4 give the results of equating Form B to Form A, Form C to Form B, and Form D to Form B under the five different methods. Raw score equivalents under the Angoff Design V methods show a strong tendency to be intermediate in value between the corresponding Tucker and Levine equivalents. Also, Design V slopes and intercepts fall between the respective Tucker and Levine slopes and intercepts.

Insert Tables 2, 3, and 4 about here

---

2. Chain equating

Tables 5, 6, and 7 show the results of equating Form A to itself around Chains 1 and 2 and equating Form B to itself around Chain 3. Figure 1 depicts these chains.

Insert Tables 5, 6, and 7 about here

Also given in these tables are estimations of equating bias, $\widehat{\text{BIAS}}$, and estimated root-mean-squared error of equating, $\widehat{\text{RMSE}}$. Bias is defined as

$$\widehat{\text{BIAS}} = \bar{X}' - \bar{X},$$

in which $\bar{X}$ is the mean raw score of a given form, and $\bar{X}'$ is the mean of the raw score equivalents obtained by equating the form to itself around a chain. Estimated root-mean-squared error, $\widehat{\text{RMSE}}$, is defined by the expression

$$\widehat{\text{RMSE}} = \sqrt{\Sigma n_i (x_i' - x_i^2)^2 / \Sigma n_i},$$

in which x_i is the i-th raw score on the base form, x_i' is its raw score equivalent, and n_i is the number of examinees who obtained the score x_i. When there are many examinees as in this study, $\widehat{\text{RMSE}}$ is calculated more easily by the expression

$$\widehat{\text{RMSE}} = \sqrt{(m-1)^2 \text{ Var } (X) + (\widehat{\text{BIAS}})^2} \ .$$

where m is the slope of the equating function $x' = mx + b$, and var(X) is the variance of X. Note that the estimated bias, $\widehat{\text{BIAS}}$, contributes to $\widehat{\text{RMSE}}$.

The departures of equivalent raw scores from original raw scores are quite similar under the Tucker, Levine, and Design V methods. In Chain 1 (Table 5), $\widehat{\text{RMSE}}$ ranges from 0.67 for the V1 method to 0.70 for the Tucker method. $\widehat{\text{BIAS}}$ of the Levine and V3 methods are somewhat lower in Chain 1 than those of the other methods. In Chain 2 (Table 6), $\widehat{\text{RMSE}}$'s of the five methods range from 0.67 (Levine and V1) to 0.73 (Tucker). As in Chain 1, the Levine and V3 methods in Chain 2 show somewhat lower estimated bias than the other three methods. In Chain 3, however, the Levine and V3 methods show somewhat higher $\widehat{\text{BIAS}}$ and $\widehat{\text{RMSE}}$ than the other methods.

In these three chains, the tendency (seen in the single-link equatings) for Design V equivalents to lie in value between corresponding Tucker and Levine equivalents appears strongly only for method V1. Method V3 produces a majority of such intermediate equivalents only in Chain 2.

3. Parametric model comparisons

Tables 8-12 start with a highly simplified and idealized equating situation and finish with one that more closely approximates those found in practice.

Table 8 gives the results of applying the equating functions of the five methods to parallel 100-item forms using a 100-item external anchor test

parallel to the forms and examinee groups that differ markedly in average performance. The Levine and V1 methods yield the identity equating function, x = y; the Tucker, V2, and V3 methods also yield slopes of 1.0, but intercepts that differ from zero respectively by 2.0, 2.0, and -2.2 raw score points.

Insert Table 8 about here

In Table 9 the populations are made equal in performance and variability. The scales of Forms X and Y are shifted by 20 points. Result: All five methods give exact equivalents.

Insert Table 9 about here

Table 10 uses parallel forms, groups of very different average ability, and an internal anchor test half the lengths of Forms X and Y. The anchor

---

Insert Table 10 about here

---

test is such that combining its items with those of a test parallel to it ("doubling it") results in a test parallel to Forms X and Y. The stipulated Form X and Form Y reliabilities of .90 and standard deviation of 20 determine the given anchor-test reliabilities, standard deviations, and anchor-form correlations. Results: The slopes of all five equating functions still equal 1.0, but notable differences appear in the intercepts. Here the Tucker and Angoff V2 intercepts are the largest and are nearly equal, whereas the Levine and V3 intercepts barely differ from zero. These results are consistent with the relatively small $\widehat{\text{BIAS}}$ and intercepts found in Chains 1 and

2 for these methods, but not with the relatively greater $\widehat{BIAS}$ and intercepts found in Chain 3.

Table 11 shows what happens when the scales of the equated forms differ by 4 raw score points and the anchor test is intermediate in difficulty between the forms. Slopes of the five equating functions are still 1.0 and the Levine method yields the most accurate results. Under these conditions the Tucker and V2 methods yield identical equating functions.

Insert Table 11 about here

Table 12 allows the two populations to have both unequal means and variabilities on the anchor test. Slopes of the five equating functions all differ; only the slope by the Levine method equals 1.0. Under this set of assumed conditions, the Levine method yields the most accurate equating.

Insert Table 12 about here

Table 13 tabulates the equatings of Table 12 for selected raw scores.

Insert Table 13 about here

Discussion

Results under the three different procedures of single-link equatings, cyclical equatings, and contrived parametric equatings show that, in the situations studied here, the Angoff Design V methods yield raw score equivalents close to those of the Tucker and Levine methods. In the cyclical or chain equatings of forms to themselves, the five methods yield comparable

estimates of bias and root-mean-squared error. In the single-link equatings, Design V raw score equivalents tend to fall between the Tucker and Levine equivalents.

However, in detail the results lack consistency among the three procedures. In the chain and parametric equatings only the method V1 equivalents fall consistently between the Tucker and Levine ones. The contrived parametric situations of Tables 10-13 show an overall advantage for the Levine method. The cyclical equatings, in contrast, fail to show such a marked advantage.

That the Levine method shows up best in the parametric situations is not surprising, for the situations are constructed in accordance with the basic assumption of the Levine method: true scores on the anchor test correlate 1.0 with true scores on at least one of the forms to be equated (Levine, 1956, p. 105). Each of the situations in Tables 8-13 involves anchor tests whose true scores correlate 1.0 with those of the forms. This condition is approximated but not fully met in the test forms used in the empirical single-link and cyclical procedures.

References

Angoff, W. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education, 508-600.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin (Eds.), Test equating. New York: Academic Press, 9-49.

Brennan, R. L., & Kolen, M. J. (1986, April). Practical issues in horizontal equating with special emphasis on the common-item nonequivalent groups design. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Levine, R. (1956). Equating the score scales of alternate forms administered to samples of different ability. (Doctoral dissertation, Syracuse University, 1955). Dissertation Abstracts International, 16, 1842A.

Appendix A (Adapted from Brennan & Kolen, 1986, April)

Summary of Equating Equations of the
Tucker and Levine Methods

General Equations for Common Item Linear Equating with Nonequivalent Populations

$$\ell(x) = \frac{\sigma_s(Y)}{\sigma_s(X)}\left[x - \mu_s(X)\right] + \mu_s(Y) \quad , \quad \text{where}$$

$$\mu_s(X) = \mu_1(X) - w_2\gamma_1\left[\mu_1(V) - \mu_2(V)\right]$$

$$\mu_s(Y) = \mu_2(Y) + w_1\gamma_2\left[\mu_1(V) - \mu_2(V)\right]$$

$$\sigma_s^2(X) = \sigma_1^2(X) - w_2\gamma_1^2\left[\sigma_1^2(V) - \sigma_2^2(V)\right] + w_1 w_2\gamma_1^2\left[\mu_1(V) - \mu_2(V)\right]^2$$

$$\sigma_s^2(Y) = \sigma_2^2(Y) + w_1\gamma_2^2\left[\sigma_1^2(V) - \sigma_2^2(V)\right] + w_1 w_2\gamma_2^2\left[\mu_1(V) - \mu_2(V)\right]^2$$

Tucker Method

$$\gamma_1 = \alpha_1(X|V) = \sigma_1(X,V)/\sigma_1^2(V)$$

$$\gamma_2 = \alpha_2(Y|V) = \sigma_2(Y,V)/\sigma_2^2(V)$$

Levine Method for Inclusive Equating Section (Using Angoff Error Estimates)

$$\gamma_1 = 1/\alpha_1(V|X) = \sigma_1^2(X)/\sigma_1(X,V)$$

$$\gamma_2 = 1/\alpha_2(V|Y) = \sigma_2^2(Y)/\sigma_2(Y,V)$$

Levine Method for Exclusive Equating Section (Using Angoff Error Estimates)

$$\gamma_1 = [\sigma_1^2(X) + \sigma_1(X,V)]/[\sigma_1^2(V) + \sigma_1(X,V)]$$

$$\gamma_2 = [\sigma_2^2(Y) + \sigma_2(Y,V)]/[\sigma_2^2(V) + \sigma_2(Y,V)]$$

Note. For use of the Tucker and Levine Methods in the present study, the test populations were given equal weight in the synthetic population: $w_1 = w_2 = 0.5$. In the above formulas, subscript 1 denotes population 1, which takes Form X, and subscript 2 denotes population 2, which takes Form Y. Thus in the notation of the Design V derivations in the Method section, $\mu(X) = \mu_1(X)$, $\mu_x(V) = \mu_1(V)$, etc.

Table 1

Examinee Groups

Certification Test Form	N	Raw Score	
		Mean	SD
A	10,450	158.47	34.91
B	22,562	162.85	35.94
C	11,272	168.08	39.31
D	11,041	156.47	36.24
E	23,450	160.55	36.25

Table 2

Equating Form B Raw Scores to Form A Raw Scores

Form B Raw Score	Form A Raw Score Equivalent				
	Tucker	Levine	Angoff V1	Angoff V2	Angoff V3
0	9.33	11.48	**10.35**	**10.17**	**10.55**
40	47.17	48.97	**48.02**	**47.79**	**48.28**
80	85.00	86.46	**85.69**	**85.42**	**86.00**
120	122.83	123.96	**123.37**	**123.04**	**123.72**
160	160.67	161.45	**161.04**	**160.67**	161.45
200	198.50	198.94	**198.72**	198.29	199.17
240	236.33	236.43	**236.39**	235.92	236.90
280	274.17	273.92	274.07	**273.54**	274.62

Equating Functions			
Tucker	A = 0.94584 B + 9.33193	Angoff V1	A = 0.94186 B + 10.34503
Levine	A = 0.93728 B + 11.48237	Angoff V2	A = 0.94064 B + 10.16534
		Angoff V3	A = 0.94309 B + 10.55395

Note. In the Angoff V1, V2, and V3 columns, the entries in boldface are those whose values fall between the corresponding Tucker and Levine raw score equivalents.

Note. The Form B examinee group was more able on the average than the Form A group. On the common-item or anchor test, Mean (B) = 56.0552, SD (B) = 13.3085; Mean (A) = 53.9865, SD(A) = 13.7240.

Table 3

Equating Form C Raw Scores to Form B Raw Scores

Form C Raw Score	Form B Raw Score Equivalent				
	Tucker	Levine	Angoff V1	Angoff V2	Angoff V2
0	-4.06	-9.48	**-6.51**	**-5.02**	**-8.05**
40	34.69	30.33	**32.71**	**33.95**	**31.42**
80	73.43	70.14	**71.93**	**72.92**	**70.90**
120	112.17	109.95	**111.15**	**111.89**	**110.37**
160	150.92	149.76	**150.38**	**150.86**	**149.85**
200	189.66	189.57	**189.60**	189.83	189.32
240	228.41	229.37	**228.82**	**228.80**	**228.80**
280	267.15	269.18	**268.04**	**267.77**	**268.27**

Equating Functions

Tucker	B = 0.96860 C-4.05897	Angoff VA	B = 0.98054 C-6.51138
Levine	B = 0.99521 C-9.47555	Angoff VB	B = 0.97426 C-5.02215
		Angoff VC	B = 0.98687 C-8.05284

Note. In the Angoff V1, V2, and V3 columns, entries in boldface are those whose values fall between the corresponding Tucker and Levine raw score equivalents.

Note. The Form B group was more able on the average than the Form C group and less variable; on the common-item or anchor test, Mean (B) = 43.0043 SD(B) = 10.5664; Mean (C) = 41.6666, SD(C) = 11.3318.

Table 4

Equating Form D Raw Scores to Form B Raw Scores

Form D Raw Score	Form B Raw Score Equivalent				
	Tucker	Levine	Angoff V1	Angoff V2	Angoff V2
0	-1.47	-6.24	**-3.59**	**-4.96**	**-2.25**
40	40.29	36.68	**38.68**	**37.69**	**39.65**
80	82.05	79.59	**80.96**	**80.34**	**81.55**
120	123.81	122.51	**123.23**	**122.99**	**123.45**
160	165.57	165.43	**165.50**	165.64	165.35
200	207.33	208.34	**207.77**	**208.29**	**207.25**
240	249.10	251.26	**250.05**	**250.94**	**249.15**
280	290.86	294.18	**292.32**	**293.59**	**291.05**

Equating Functions

Tucker	B = 1.04403 D -1.47267	Angoff V1	B = 1.05683 D -3.59109
Levine	B = 1.07292 D -6.24012	Angoff V2	B = 1.06625 D -4.95872
		Angoff V3	B = 1.04750 D -2.24784

Note. In the Angoff V1, V2, and V3 columns, the entries in boldface are those whose values fall between the corresponding Tucker and Levine raw score equivalents.

Note. The Form B examinee group was slightly more able on the average and less variable than the Form D group. On the common-item or anchor test, Mean (B) = 42.4121, SD(B) = 10.4474; Mean (D) = 42.0973, SD(D) = 11.1320.

Table 5

Equivalent Raw Scores Obtained Under Use of Tucker Method, Levine Method, and Angoff Designs V1, V2, and V3, Equating Chain 1: Form A → Form C → Form B → Form A

Original Form A Raw Score	Raw Score Equivalent on Form A as Old Form				
	Tucker Method[a]	Levine Method[b]	Angoff V1	Angoff V2	Angoff V3
0	-0.49	-1.76	**-1.04**	-0.35#	-1.76
40	39.80	38.80	**39.37**	39.90#	**38.81**
80	80.09#	79.37	**79.77**	80.16	79.37
120	120.37	119.93	**120.18**	120.41	119.94#
160	160.66	160.50#	**160.59**	160.66	160.50#
200	200.95	201.06	**201.00**	200.91#	201.07
240	241.23	241.63	**241.40**	241.16#	241.64
280	281.52	282.19	**281.81**	281.41#	282.20
$\widehat{\text{BIAS}}$[c]	0.65	0.47	0.57	0.65	0.48
$\widehat{\text{RMSE}}$[d]	0.70	0.68	0.67	0.69	0.69

Final Equating Functions

Tucker $A' = 1.00717A - 0.48616$ Angoff V1 $A' = 1.01019\ A - 1.04152$

Levine $A' = 1.01411A - 1.76131$ Angoff V2 $A' = 1.00627\ A - 0.34571$

Angoff V3 $A' = 1.01415\ A - 1.75845$

Note. In the Angoff V1, V2, and V3 columns, the entries in boldface are those whose values fall between the corresponding Tucker and Levine equivalents. The symbol "#" denotes equivalents closest in value to corresponding original raw scores.

Note. The original raw scores have a mean of 158.47 and a sample standard deviation of 35.85.

[a]Tucker Equally Reliable Tests method with equally weighted old and new test groups.

[b]Levine Equally Reliable Tests method with equally weighted old and new test groups.

[c]Estimated bias. See pp. 10-11.

[d]Estimated root-mean-squared error. See pp. 10-11.

Table 6

Equivalent Raw Scores Obtained Under Use of Tucker Method, Levine Method, and Angoff Designs V1, V2 and V3, Equating Chain 2: Form A → Form D → Form B → Form A

Original Form A Raw Score	Raw Score Equivalent on Form A as Old Form				
	Tucker Method[a]	Levine Method[b]	Angoff V1	Angoff V2	Angoff V3
0	-0.29	-1.95	**1.00**	-0.27#	**-1.75**
40	39.96#	38.64	**39.40**	39.96#	**38.81**
80	80.21	79.24	**79.79**	80.20#	**79.37**
120	120.46	119.83	**120.19**	**120.43**	119.93#
160	160.71	160.43#	**160.59**	**160.67**	**160.49**
200	200.96	201.03	**200.99**	200.90#	201.05
240	241.21	241.62	**241.38**	241.14#	**241.61**
280	281.46	282.22	**281.78**	281.37#	**282.17**
$\widehat{\text{BIAS}}$[c]	0.70	0.41	0.57	0.66	0.47
$\widehat{\text{RMSE}}$[d]	0.73	0.67	0.67	0.69	0.69

Final Equating Functions

Tucker $A' = 1.00623\ A - 0.28770$ Angoff V1 $A' = 1.00992\ A - 0.99861$

Levine $A' = 1.01490\ A - 1.95401$ Angoff V2 $A' = 1.00587\ A - 0.27169$

Angoff V3 $A' = 1.01401\ A - 1.74928$

Note. In the Angoff V1, V2, and V3 columns, the entries in boldface are those whose values fall between the corresponding Tucker and Levine equivalents. The symbol "#" denotes equivalents closest in value to corresponding original raw scores.

Note. The original raw scores have a mean of 158.47 and a sample standard deviation of 35.85.

[a]Tucker Equally Reliable method with equally weighted old and new test groups.

[b]Levine Equally Reliable method with equally weighted old and new test groups.

[c]Estimated bias. See pp. 10-11.

[d]Estimated root-mean-squared error. See pp. 10-11.

Table 7

Equivalent Raw Scores Obtained Under Use of Tucker Method, Levine Method, and Angoff Designs V1, V2 and V3, Equating Chain 3: Form B → Form E → Form C → Form B

Original Form B Raw Score	Raw Score Equivalent on Form B as Old Form				
	Tucker Method[a]	Levine Method[b]	Angoff V1	Angoff V2	Angoff V3
0	2.65	1.80	**2.34**	3.27	1.43#
40	43.04	42.62	**42.90**	43.51	42.53#
80	83.42	83.43	83.47	83.75	83.22#
120	123.81#	124.25	**124.04**	**123.99**	**124.12**
160	164.19#	165.06	**164.60**	**164.23**	**165.02**
200	204.58	205.88	**205.17**	204.47#	205.91
240	244.96	246.70	**245.74**	244.71#	246.81
280	285.34	287.51	**286.31**	284.96#	287.70
$\widehat{\text{BIAS}}$[c]	4.22	5.12	4.64	4.25	5.08
$\widehat{\text{RMSE}}$[d]	4.23	5.17	4.67	4.25	5.14

Final Equating Functions

Tucker $B' = 1.00961\ B + 2.65313$

Levine $B' = 1.02039\ B + 1.80208$

Angoff VA $B' = 1.01418\ B + 2.33513$

Angoff VB $B' = 1.00603\ B + 3.26677$

Angoff VC $B' = 1.02240\ B + 1.43098$

Note. In the Angoff V1, V2, and V3 columns, the entries in boldface are those whose values fall between the corresponding Tucker and Levine equivalents. The symbol "#" denotes equivalents closest in value to corresponding original raw scores.

Note. The original raw scores have a mean of 162.85 and a sample standard deviation of 35.94.

[a]Tucker Equally Reliable method with equally weighted old and new test groups.

[b]Levine Equally Reliable method with equally weighted old and new test groups.

[c]Estimated bias. See pp. 10-11.

[d]Estimated root-mean-squared error. See pp. 10-11.

Table 8

Parametric Equating of Forms X and Y
Using Different Linear Methods

Forms: Parallel
Populations: Very different means, equal variability
Anchor Test: Parallel to the forms, external

	Old Test X	New Test Y	Anchor Test V	
			Population X	Population Y
No. of Items	100	100	100	100
Mean	60	40	60	40
SD	20	20	20	20
Reliability	.90	.90	.90	.90

$\mu(X) = 60$ $\mu_x(V) = 60$ $\rho(Y,V) = .90$ $\sigma_y(V) = 20$ $\sigma(Y) = 20$

$\mu(Y) = 40$ $\mu_y(V) = 40$ $\rho(X,V) = .90$ $\sigma_x(V) = 20$ $\sigma(X) = 20$

Method	Equating Function	
	Slope	Intercept
Tucker	1.00000	2.00000
Levine	1.00000	0.00000
V1	1.00000	0.00000
V2	1.00000	2.00000
V3	1.00000	-2.22222

Table 9

Parametric Equating of Forms X and Y
Using Different Linear Methods

Forms: Congeneric shifted: X = Y + 20, equally reliable
Populations: Equivalent
Anchor Test: Parallel to the forms, external

	Old Test X	New Test Y	Anchor Test V	
			Population X	Population Y
No. of Items	100	100	100	100
Mean	60	40	60	60
SD	20	20	20	20
Reliability	.90	.90	.90	.90

$\mu(X) = 60$ $\mu_x(V) = 60$ $\rho(Y,V) = .90$ $\sigma_y(V) = 20$ $\sigma(Y) = 20$
$\mu(Y) = 40$ $\mu_y(V) = 60$ $\rho(X,V) = .90$ $\sigma_x(V) = 20$ $\sigma(X) = 20$

Method	Equating Function	
	Slope	Intercept
Tucker	1.00000	20.00000
Levine	1.00000	20.00000
V1	1.00000	20.00000
V2	1.00000	20.00000
V3	1.00000	20.00000

Table 10

Parametric Equating of Forms X and Y
Using Different Linear Methods

Forms: Parallel
Populations: Very different means, equal variability
Anchor Test: Congeneric to both forms, half length of forms, internal

	Old Test X	New Test Y	Anchor Test V	
			Population X	Population Y
No. of Items	100	100	50	50
Mean	60	40	30	20
SD	20	20	10.48811	10.48811
Reliability	.90000	.90000	0.81818	0.81818

$\mu(X) = 60$ $\mu_x(V) = 30$ $\rho(Y,V) = 0.95346$ $\sigma_y(V) = 10.48811$

$\sigma(Y) = 20$ $\mu(Y) = 40$ $\mu_y(V) = 20$ $\rho(X,V) = 0.95346$

$\sigma_x(V) = 10.48811$ $\sigma(X) = 20$

Method	Equating Function	
	Slope	Intercept
Tucker	1.00000	1.81820
Levine	1.00000	0.00006
V1	1.00000	0.93079
V2	1.00000	1.81827
V3	1.00000	-0.00001

BEST COPY AVAILABLE

Table 11

Parametric Equating of Forms X and Y
Using Different Linear Methods

Forms: Congeneric shifted: X = Y + 4, equally reliable
Populations: Nonequivalent: unequal means, equal variabilities
Anchor Test: Congeneric to Forms X and Y, intermediate in difficulty between the forms, half length of forms, external

	Old Test X	New Test Y	Anchor Test V	
			Population X	Population Y
No. of Items	100	100	50	50
Mean	50	48	24	25
SD	20	20	10.48811	10.48811
Reliability	.90	.90	0.81818	0.81818

$\mu(X) = 50$ $\mu_x(V) = 24$ $\rho(Y,V) = 0.85812$ $\sigma_y(V) = 10.48811$

$\sigma(Y) = 20$ $\mu(Y) = 48$ $\mu_y(V) = 25$ $\rho(X,V) = 0.85812$

$\sigma_x(V) = 10.48811$ $\sigma(X) = 20$

Method	Equating Function	
	Slope	Intercept
Tucker	1.00000	3.63637
Levine	1.00000	4.00001
V1	1.00000	3.90692
V2	1.00000	3.63637
V3	1.00000	4.22221

Note. A population scoring 25 on the anchor averages 52 on Form X (the easier form) and 48 on Form Y.

Note. $X_T = 2V_T + 2$, $Y_T = 2V_T - 2$, where subscript T denotes true score.

BEST COPY AVAILABLE BEST COPY AVAILABLE

Table 12

Parametric Equating of Forms X and Y
Using Different Linear Methods

Forms: Congeneric shifted: X = Y + 4, equally reliable
Populations: Nonequivalent: unequal means, unequal variabilities
Anchor Test: Congeneric to Forms X and Y, intermediate in difficulty between the forms, half length of forms, external

	Old Test X	New Test Y	Anchor Test V	
			Population X	Population Y
No. of Items	100	100	50	50
Mean	50	48	24	25
SD	18.97366	20	10.00000	10.48811
Reliability	.88889	.90	0.80000	0.81818

$\mu(X) = 50$ $\mu_x(V) = 24$ $\rho(Y,V) = 0.85812$ $\sigma_y(V) = 10.48811$

$\sigma(Y) = 20$ $\mu(Y) = 48$ $\mu_y(V) = 25$ $\rho(X,V) = 0.84328$

$\sigma_x(V) = 10$ $\sigma(X) = 18.97366$

Method	Equating Function	
	Slope	Intercept
Tucker	0.98197	4.46890
Levine	1.00000	3.99989
V1	0.99499	4.13789
V2	0.97778	4.66647
V3	1.01250	3.65003

Note. A population averaging 25 on the anchor test V would have a mean of 48 on Form Y and a mean of 52 on Form X.

Note. Computed reliability of Form X differs from that of Form Y because the populations taking these forms have different standard deviations.

BEST COPY AVAILABLE

Table 13

Simulated Equating of Congeneric Forms Via an
External Anchor and Populations of Unequal Means and Variabilities
Using the Equating Functions of Table 12

Form Y	Equivalent Form X Raw Score Under Five Equating Methods				
Raw Score	Tucker[a]	Levine[b]	Angoff V1	Angoff V2	Angoff V3
0	4.47	4.00	**4.14**	4.67	3.65
10	14.29	14.00	**14.09**	14.44	13.78
20	24.11	24.00	**24.04**	24.22	23.90
30	33.93	34.00	**33.99**	34.00	34.03
40	43.75	44.00	**43.94**	**43.78**	44.15
50	53.57	54.00	**53.89**	53.56	**54.28**
60	63.39	64.00	**63.84**	63.33	**64.40**
70	73.21	74.00	**73.79**	73.11	74.53
80	83.03	84.00	**83.74**	82.89	84.65
90	92.85	94.00	**93.69**	92.67	94.78
100	102.67	104.00	**103.64**	102.44	104.90
Slope	0.98197	1.00000	0.99499	0.97778	1.01250
Intercept	4.46890	3.99989	4.13789	4.66647	3.65003
BIAS[c]	-0.40	0.00	-0.10	-0.40	0.25
RMSE[d]	0.54	0.00	0.14	0.60	0.35

Note. In the Angoff V1, V2, and V3 columns, entries in boldface are those whose values fall between the corresponding Tucker and Levine raw score equivalents.

[a]Tucker Method for equally reliable tests. Populations of both forms were equally weighted in the synthetic population.

[b]Levine Method for equally reliable tests. Populations of both forms were equally weighted in the synthetic population.

[c]See pp. 10-11.

[d]Root-mean-squared error. See pp. 10-11.

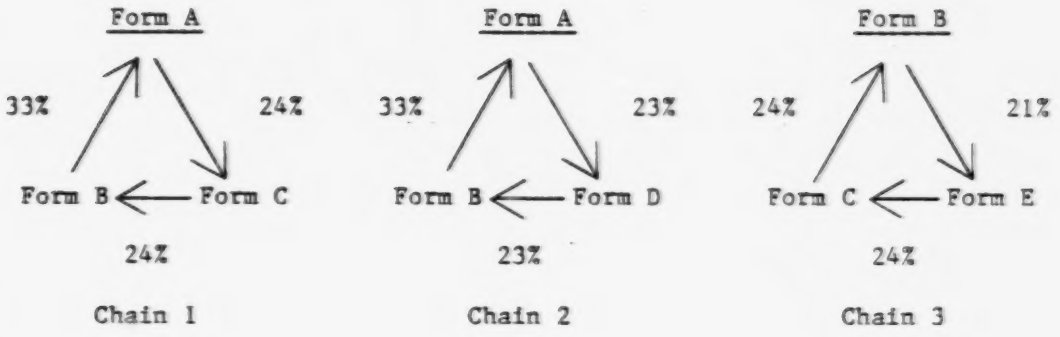

Figure 1. Chains used to compare accuracy of equating under use of Tucker, Levine, and Angoff Design V equating methods. Arrows indicate equating links. A percentage adjacent to an arrow gives the number of equating items as a percentage of the number of items in each of the equated forms.